IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON ADVANCED BIG DATA ANALYSIS FOR VEHICULAR SOCIAL NETWORKS

Received December 25, 2018, accepted January 6, 2019, date of publication January 11, 2019, date of current version April 13, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2892469

Infrared Multi-Pedestrian Tracking in Vertical View via Siamese Convolution Network

GUOJIANG SHEN[1], LINFENG ZHU[1], JIHAN LOU[2], SI SHEN[1], ZHI LIU[1], AND LONGFENG TANG[1]

[1]Computer Intelligence System Institute, Zhejiang University of Technology, Hangzhou 310023, China
[2]Zhejiang Supcon Information Co., Ltd., Hangzhou 310052, China

Corresponding author: Zhi Liu (lzhi@zjut.edu.cn)

This work was supported by the Scientific Research Project of Education Department of Zhejiang under Grant Y201840830.

ABSTRACT Target tracking has become one of the research hotspots in the field of computer vision in recent years. In this paper, a new intelligent algorithm of infrared multi-pedestrian tracking in vertical view is proposed. In the algorithm, the pedestrians in the infrared image can be quickly detected and located with the method of the Faster Regions with CNN features (RCNN) and then are tracked with the improved Siamese network. The tracking method based on Siamese network transforms the tracking problem into a similarity verification problem and evaluates the similarity score between new frame feature and target frame feature by convolution network. The candidate region with the highest score is considered as the current position of the target. In this paper, the Siamese network is combined with Faster RCNN for multi-pedestrian tracking. In addition, the tracking results of adjacent frames are introduced into the similarity evaluation of current frames to improve the tracking accuracy when the pedestrian posture changes. The experimental results show that the algorithm has good robustness and tracking result and achieves competitive performance.

INDEX TERMS Computer vision, Siamese network, infrared detection, pedestrian tracking, convolution network.

I. INTRODUCTION

Target tracking is one of the most concerned subjects in the field of computer vision. With the increasing application of artificial intelligent in daily life, it is particularly important to use target tracking technology to track pedestrians, vehicles or other objects on the road and provide real-time and accurate information [1]–[4]. In recent years, various effective tracking algorithms [5]–[7] have been proposed constantly and shown excellent performance in some data sets. Among them, the pedestrians can be tracked, the pedestrian motion can be detected, and pedestrian numbers in the area can be identified. Compared with visible light, the pedestrian tracking under infrared light is insensitive to the illumination changes and can be operated in all-weather. In addition, infrared images consist of only grayscale information, which can protect sensitive information in some private scenes. However, infrared images have some limitations, such as less color information and low resolution.

In view of the excellent performance of deep convolutional neural network in feature extraction, it has been the state of art of target tracking [8], [9]. Some researchers have combined convolution features with traditional tracking methods and achieved quite good results [6], [7]. So as to improve the tracking effect, some researchers have proposed to transform the target tracking problem into a similarity verification problem [10], [12], [14]. In addition, Bertinetto L proposed a fully convolution Siamese network (SiamFC), which trained two same full convolution networks, one for extracting target features and the other for analyzing candidate region features. By comparing the two set of features, the target location could be found [14].

In view of the outstanding performance of SiamFC in target tracking, this paper focus on the infrared pedestrian tracking method in vertical view by SiamFC in which the following three problems need to be solved. Firstly, SiamFC lacks the updating of target model. In the whole tracking process, only the first frame is chosen as the target model (the target needs

to be pre-marked), without considering the change of posture during the motion. Secondly, SiamFC only tracks the single target, so the algorithm needs to be improved when it is applied to multi-object tracking. Finally, since the image used in this paper is an infrared pedestrian image from the vertical view, there are hardly any public data sets.

To solve these problems, the following improvements have been made in this paper. Firstly, the tracking results of the previous frame are added into the similarity evaluation of the current frame, and the target model is indirectly updated. Secondly, by Faster-RCNN, all the pedestrians included in the image are detected and the status (being track, to be tracked or no longer tracked) of each pedestrian is identified. Then, according to the statuses of each pedestrian, multi-object tracking is realized through SiamFC. Finally, in order to overcome the shortage of training data, video data recorded in actual scene is utilized as Faster-RCNN's training data. For Siamese network, many large visible video data sets are available, such as OTB, VOT and so on.

The remaining of this paper is organized as follow. Reviews of pedestrian detection model on Faster-RCNN and object tracking model on Siamese network are introduced in Section II. Methodology is depicted in Section III, and experimental details and results are provided in Section IV. Finally, the conclusions and discussion are presented in Section V.

II. RELATED WORK

A. PEDESTRIAN DETECTION BASED ON FASTER-RCNN

The research content of this paper is to realize the tracking of the specific object – pedestrians. In the actual situation, pedestrians cannot be pre-marked in the video image, so how to detect the pedestrians in each frame is the premise for tracking.

Nowadays, the target detection and recognition based on deep learning becomes the mainstream and many methods have emerged, Such as R-CNN [24] and its improved version like Fast-RCNN [23], Faster-RCNN [17], R-FCN [25], SSD [18] and YOLO [19]. RCNN was proposed by Girshick in 2014 and known for its high accuracy on object detection. However, this method has heavy computational burden. Then Girshick proposed Fast-RCNN to improve the real-time performance of computing, but there is a bottleneck of region proposal computation. Based on Fast-RCNN, the Faster-RCNN was proposed in 2016. It consists of two networks, one is the Region Proposal Network (RPN) which is a fully convolutional network to generate proposal region and pass it to next part, the other is a detection network for target detecting and classifying. The key point of Faster-RCNN is to share full-image convolutional features between RPN and detection network, so the region proposal almost cost-free.

Since Faster-RCNN was proposed, it has been one of the mainstream frameworks in the field of target detection. For general detection problems (such as pedestrian detection, vehicle detection, and text detection), Faster-RCNN can achieve good results both in accuracy and speed.

B. PEDESTRIAN TRACKING BASED ON SIAMESE NETWORK

The target tracking problem can be transformed into a similarity problem. By comparing the target image with search area, the region with the highest similarity can be regarded as the location of target. The advantage of this method is that the pre-training network can be used to track targets, and the real-time tracking can be improved.

Siamese network is very suitable for solving similarity learning problem. It consists of two branches, one is to extract the features of the target, and the other is to divide the search image into several candidate sub-regions, extract the features of these sub-regions, and then evaluate the similarity between the target and sub-regions. The region with the highest score is the track result. An excellent method based on Siamese network was proposed by Bertinetto *et al.* [14], called full convolution Siamese network (SiamFC). This method uses the full convolution network, and its advantage is that the larger search image can be selected as the input and the similarity function will calculate all transformed sub-windows within the search image in one evaluation. This method can track a given single target in real time and has good precision.

The excellent performance of SiamFC has attracted wide attention in target tracking and its follow-up work [15], [16], [22], [26], [27] . SiamRPN improved the speed and accuracy of tracking algorithm by using the candidate sub-regions recommended by RPN [15]. SA-Siam constructed a dual Siamese network to learn semantic features and appearance features, and combine these two features to improve tracking accuracy [12].

III. INFRARED MULTI-PEDESTRIAN TRACKING FROM VERTICAL VIEW

A. FASTER-RCNN FOR PEDESTRIAN DETECTION

In this paper, Faster-RCNN is used to locate pedestrians in video, and its algorithm framework is shown in figure 1. The convolutional network adopts VGG16 [20], and its training set is extracted from the infrared pedestrian video recorded in the actual scenes.

As show in figure 1, the convolution network is utilized to obtain the feature map of the image and share them to the regional proposal network (RPN), RPN uses softmax to determine whether each anchor point belongs to the foreground or background. ROI Pooling layer generates proposal feature maps based on feature map and proposal region which are used to determine the target category, and then passes them to the full connection layer. In the end, softmax classifier discriminates whether each proposal region is pedestrian, and the final target position is obtained by bounding box regression.

B. SiameseFC BASED PEDESTRIAN TRACKING

In this paper, SiameseFC is used to track pedestrian, and its network framework is shown in figure 2. The whole process

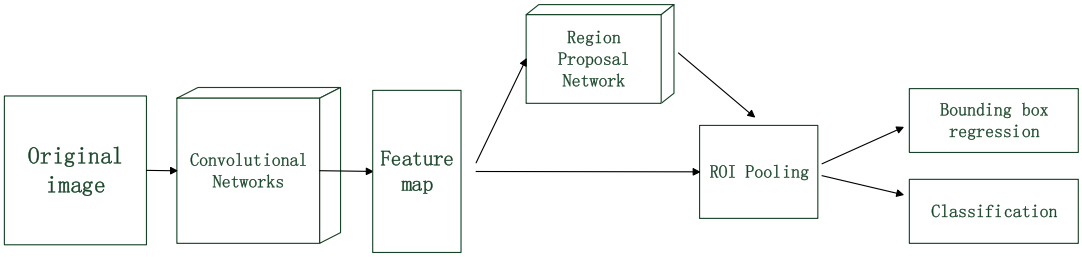

FIGURE 1. Faster RCNN framework.

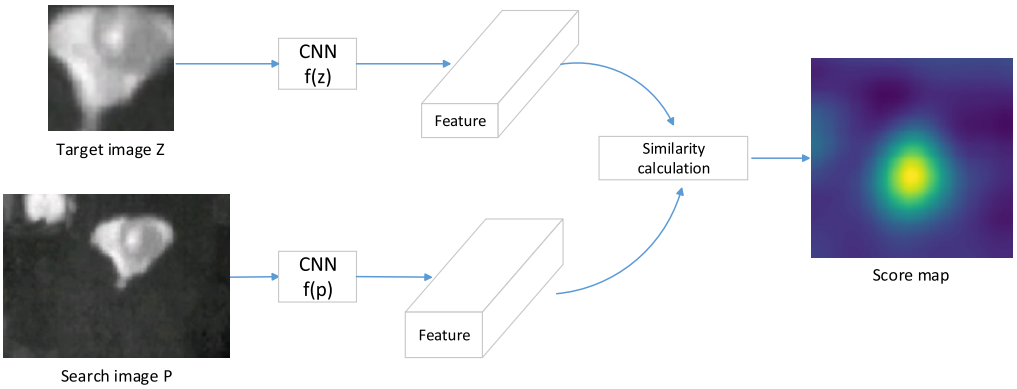

FIGURE 2. Siamese network framework.

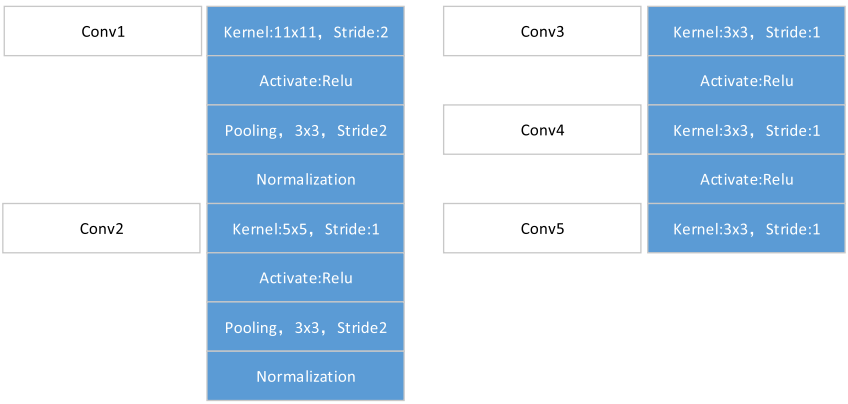

FIGURE 3. CNN network architecture.

is expressed as follows:

$$g(z, p) = \text{corr}(f(z), f(p)). \quad (1)$$

where z represents the image of the target model, p stands for the searching image, $g(z, p)$ denotes the similarity between z and p, $f(\cdot)$ denotes the feature extracted by convolution models, and $\text{corr}(\cdot)$ is a function that measures similarity, such as Euclidean distance.

The convolution feature extraction layer represented as $f(\cdot)$ has a similar structure to that of AlexNet [21], but there is no the final full connection layer. As depicted in figure 3, the whole CNN has five convolutional layers. There is a maximum pooling layer behind conv1 and conv2, and except conv5 there are a ReLU non-linear activation layer behind each convolutional layer.

C. SIAMESE CONVOLUTION NETWORK TRAINING

The training samples are taken from the large marked video dataset in ImageNet. To ensure the same size of training image, all infrared video images should be scaled according to the scale s. The target to be tracked and the search area centered on the target to be tracked are cut out by each frame image in a fixed size. Then, a pair of training sample is formed by randomly selecting a target image and its corresponding search region image (the interval between the two is no more than T frames) in the same video. If the size of the labeled

box is $w \times h$, and the size of the target image to be cut is A^2, then there is

$$A^2 = s\,(w + 2p) \times s\,(h + 2p)\,. \tag{2}$$

where p represents the width of the edge and $p = (w + h)/4$, and s represents the factor of image scaling.

During training, the search image of each pair of training sample produces multiple candidate regions, which were divided into positive or negative candidate regions, as defined follows:

$$y\,[u] = \begin{cases} +1 & \text{if } k\,\|u - c\| \le R \\ -1 & \text{else.} \end{cases} \tag{3}$$

where u represents candidate region, k denotes the network stride, c stands target center, and R indicates the preset radius. That is to say, if the candidate region is within the radius R of the target center, it is recognized as positive candidate region, otherwise it is a negative candidate region.

Each candidate region corresponds to a sample. And its score can be seen as the probability that it is a positive or negative sample. So, its logical loss can be expressed as:

$$\varepsilon\,(y, v) = \log\,(1 + \exp\,(-yv))\,. \tag{4}$$

where v represents the actual output of a candidate region, and $y \in \{+1, -1\}$ denotes that the candidate region is positive or negative. Therefore, the loss function of the entire network can be expressed as:

$$L\,(y, v) = l\frac{1}{|D|}\sum_{u \in D} \varepsilon\,(y\,[u]\,, v\,[u])\,. \tag{5}$$

where $D \in R^2$ represents the score map containing all candidate regions of a sample pair.

D. TRACKING TEMPLATE UPDATING

SiamFC [14], SINT [22] and other methods generally regard the target image in the first frame as the tracking template, and in the subsequent tracking process, the target template is not updated. They consider the target image in the first frame to be the most important and credible data. During the similarity evaluation, all candidate regions in the search image are compared with the target template, and the candidate regions with the highest similarity score are returned as the final position of the target. This process can be expressed as:

$$p' = \underset{p_k \in p}{\arg\ \max}\ g\,(z, p_k) \tag{6}$$

where z is the initially selected target image, p' represents the candidate region with the highest similarity to z, p denotes all candidate regions, p_k represents the k candidate region, $g\,(z, p_k)$ denotes the similarity score between z and p_k.

However, it is quite possible for the target to produce deformation, occlusion and other variation during the moving, which leads to significant difference between original template and the subsequent image. As a result, the bias accumulated gradually will reduce the accuracy of the follow-up tracking. In general, pedestrian changes are very small in adjacent frame. When the difference between the pedestrian image in the current frame and the initial image is large, the current tracking result can be corrected by using the tracking result of the previous frame. Therefore, this paper fuses the previous frame tracking results into the similarity evaluation, and transforms equation (6) into

$$p' = \underset{p_k \in p}{\arg\ \max}\left(\lambda g\,(z, p_k) + (1 - \lambda)\,g\left(p'_{t-1}, p_k\right)\right) \tag{7}$$

where p'_{t-1} represents the candidate region with the highest similarity to the target z in the previous frame, and λ is a weight coefficient .The whole similarity score has two parts: one is the similarity score between the first frame target and the candidate region of the current frame, and the other is score between the highest similarity candidate region of the previous frame and the candidate region of the current frame.

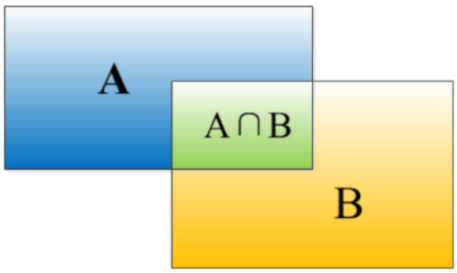

FIGURE 4. IOU Schematic diagram.

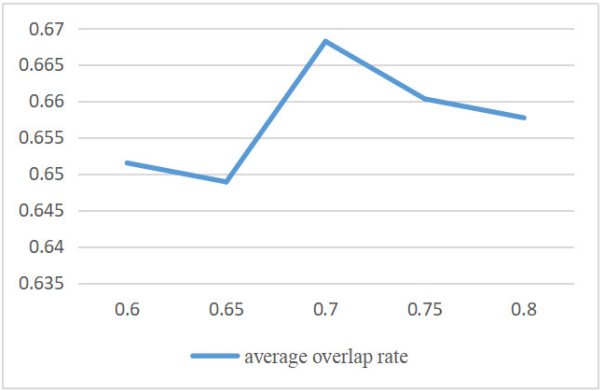

FIGURE 5. Overlap rate with different λ.

E. MULTI-PEDESTRIAN TRACKING

I_y The following sets are defined firstly: the set I is a set of pedestrians detected in the current frame, represents the pedestrian y in set I which contains the position coordinates of the pedestrian y and the image of the current frame, The set H is a set of pedestrians which need to be tracked in the current frame, H_x represents the pedestrian x in set H which includes the historical position coordinates and the corresponding image of x from entering the monitoring area to the current frame, The set $H' \subseteq H$ is the tracking results

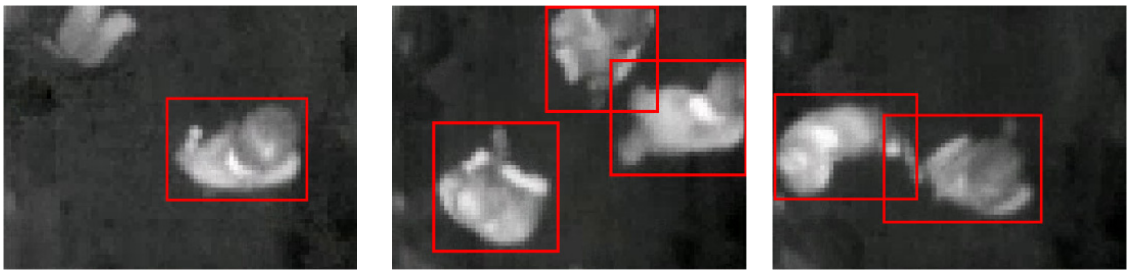

FIGURE 6. Infrared pedestrian detection results.

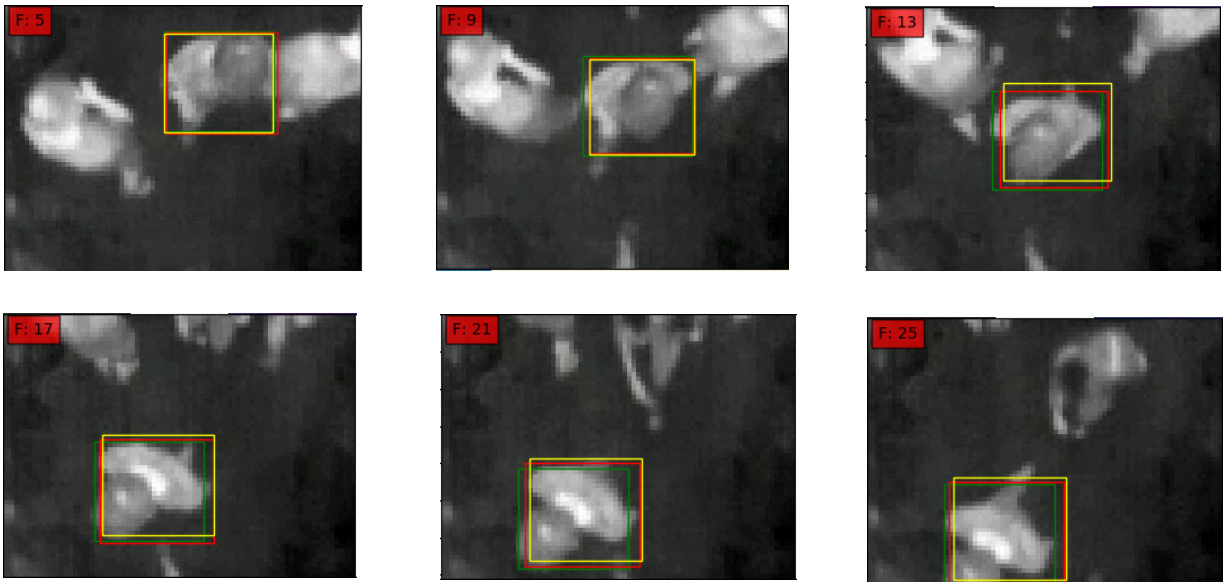

FIGURE 7. Pedestrian tracking results.

of the pedestrians in the current frame, and H'_x represents the tracking result of the pedestrian x in the current frame.

The multi-pedestrian tracking process is as follows:

1. The Faster-RCNN is used to detect pedestrians in the current frame, and the set I is obtained.

$$I = F(Z). \tag{8}$$

where $F(\bullet)$ represents the pedestrian detection operation of Faster-RCNN and Z is the current frame image.

In the current frame, the pedestrians in the set H is tracked by SiamFC, and the set H' is obtained.

$$H'_x = G(Z, H_x), \quad H'_x \in H', \ H_x \in H \tag{9}$$

where $F(\bullet, \bullet)$ represents the pedestrian tracking operation of SiamFC.

3. The set H is updated according to the change of set I and H'.

In order to update the set H, IOU(Intersection over Union) is defined to calculate the position overlap rate between the pedestrian A in the set I and the pedestrian B in the set H'. If $IOU(A, B)$ is greater than the threshold T, A and B can be regarded as the same pedestrian, otherwise it is not. In Fig. 4 the blue portion indicates the area occupied by A in the image, the yellow portion denotes the area occupied by B in the image, and the green portion represents the part of A and B intersecting. $IOU(A, B)$ equals the intersection of A and B divided by the union of A and B.

$$IOU(A, B) = \frac{A \cap B}{A \cup B} \tag{10}$$

In addition to IOU, the boundary of the monitoring area is defined which is a rectangle slightly smaller than the image. The purpose is to ensure that the detected pedestrian contour is complete and the accuracy of the tracking is guaranteed.

The update to the set H is divided into two steps:

(1) Identifying the newly added pedestrians. The IOU between the pedestrian P in the set I and all pedestrians in set H' are calculating. If there is H'_x such that the maximum value of $IOU(I_P, H'_x)$ is less than the threshold T, and the pedestrian P is within the boundary of the monitoring area, then pedestrian P is considered to be a newly pedestrian. At this time, the relevant information of pedestrian I_P is added to the set H.

(2) Judging the pedestrian who has just left. For the pedestrian q in set H, if the tracking result H'_q in set H' is not

within the boundary of the monitoring area, it indicates that the pedestrian q has left, and its corresponding information H_q should be removed from set H. Otherwise, the position coordinates in H'_q are updated to H_q.

4. After the update of set H is completed, the new set H is used for pedestrian tracking in the next frame.

IV. EXPERIMENT

A. MODEL TRAINING

In this paper, GeForce gtx1060 6G GPU is used to experiment in windows python3.5 environment.

The training set of Faster-RCNN is about 5000 frames of pedestrian images recorded by a fixed infrared camera in an aisle. The real boundary box of the target are manually labeled, and the marked data was saved into a text file. The open source tensorflow framework, and VGG16 network model are used. The number of iterations for training is set to 40,000, and the learning rate is 0.001.

SiamFC can track any pre-marked single target, so it is not necessary to collect special video to train it. The training data in this article comes from ImageNet's large video set (ILSVRC2015). SiamFC training consists of more than 50 epochs, each of which contains 50,000 pairs of samples, and the gradient of each iteration is calculated using 8 as the min-batch. The learning rate for each epoch is from 10^{-2} to 10^{-5}.

The parameter λ in the formula (7) represents the weight between the initial model and the optimal candidate of the previous frame in the similarity evaluation of the current frame. It is found by experiments that the overlap rate between the trailing box and the actual labeling box is different with different λ values. As shown in figure 5, tracking is best when $\lambda = 0.7$.

B. EXPERIMENT RESULT

The trained Faster-RCNN is used to detect the non-training data set of the infrared pedestrian in the vertical view. Some detection results are shown in Fig. 6. The experimental results show that the excellent detection results can be obtained for the images with clear pedestrian boundaries and complete pedestrian. However, if pedestrians are quite close to each other, or only a part of the body appear in the image, the accuracy will drop. A large number of tests show that the detection accuracy of Faster-RCNN is about 91.2% and the detection speed is about 0.015s per frame. The detection performance basically meets the requirements of practical application.

The trained SiamFC network is used in the infrared pedestrians tracking experiment. The experimental results show that the tracking accuracy of the same pedestrian decreases gradually with the increase of tracking time. Here, the accuracy of pedestrian tracking is expressed by the overlap rate between the pedestrian tracking box and the pre-marked box. In the early stage of the tracking, the pedestrian's posture is similar to those of the preset pedestrian model, so the overlap rate is between 0.85 and 0.95. With the change of walking posture, the overlap rate will drop to about 0.63. Experimental results show that the overlap rate of the algorithm proposed in this paper is increased from 0.63 to 0.67 for the same target in the same video, and the accuracy has been improved to a certain improvement. In Figure 7, the green border is the actual box of the pedestrians marked manually, the yellow is the SiamFC tracking result, and the red is the improved tracking result. It can be seen from the figure that the pedestrian posture is barely changed in the early stage of tracking, and the three boxes almost overlap completely. With the change of pedestrian posture, the red and the yellow boxes begin to deviate. However, compared with the yellow box, the red is closer to the real box.

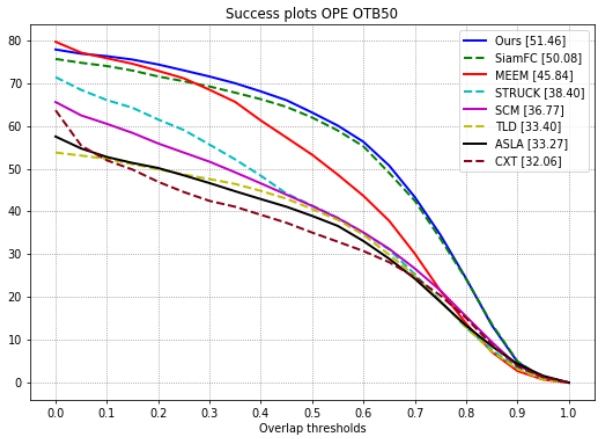

FIGURE 8. OTB result.

In addition, the improved algorithm is compared with the original algorithm and other mainstream tracking algorithms on the OTB50. The tracking success rate (AUC) is shown in Figure 8. In the Figure, the abscissa represents the threshold of the overlap rate between the tracking box and the real pre-marked box, and the ordinate represents the tracking success rate. When the overlap rate is greater than this threshold during the tracking process, the tracking is considered to be successful. It can be seen from figure 8 that the tracking success rate of SiamFC algorithm is ahead of that of most other algorithms, and it can maintain a relatively good success rate with the increase of threshold value. The proposed tracking algorithm in this paper improves the similarity evaluation process on the basis of SiamFC and increases the tracking success rate by about 3% compared with the original algorithm.

When pedestrians enter or leave the image boundary, only part of the body appears in the image, which is not conducive to pedestrian detection and tracking. Therefore, a monitoring area boundary is set up in this paper. The pedestrian will only be tracked when the detected pedestrian center is within this boundary. The multi-person pedestrian tracking effect is shown in Figure 9.

Figure 9 captures 6 frames from a consecutive 22-framevideo. In each frame, different colors of the box

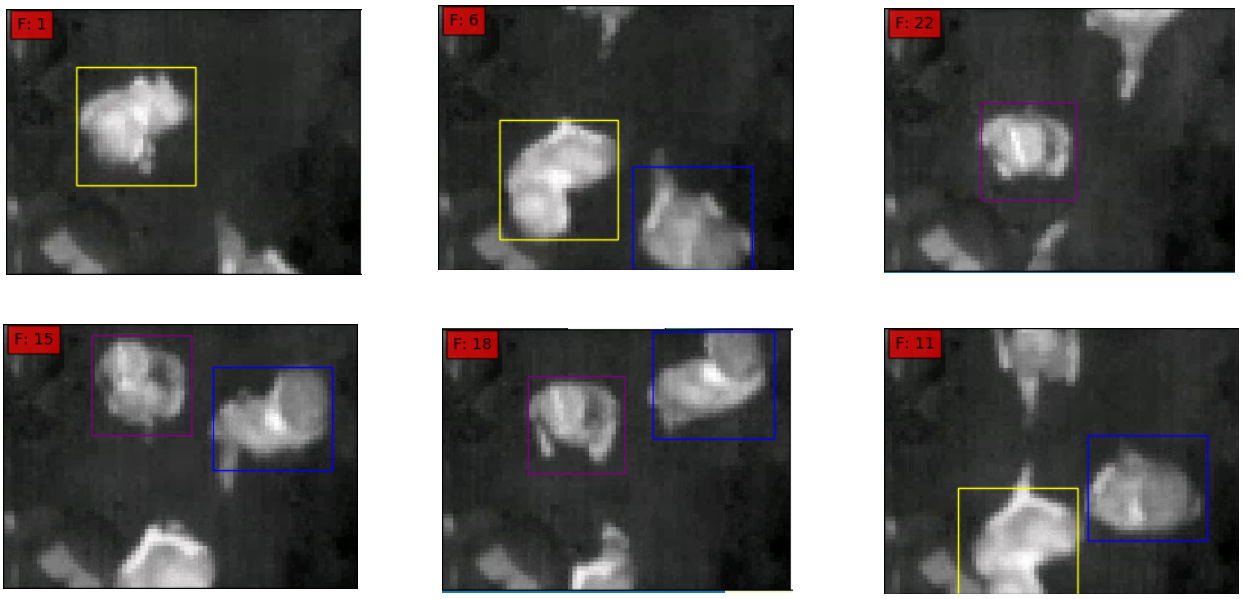

FIGURE 9. Multi-pedestrian tracking process.

represent different pedestrians. The pedestrians from entering to leaving the monitoring area are represented by only one color. In Figure 9, the yellow box from frame1 to frame 11, the blue box from frame 6 to frame 18, and the purple box from frame 15 to frame 22 respectively indicate the process of tracking different pedestrian trajectories. It can be seen that infrared multi-pedestrian tracking is effective and meets the requirements of practical applications.

V. CONCLUSION

Infrared multi-pedestrian tracking in vertical view have broad application prospects. It can be used to calculate the flow of pedestrians on the sidewalk and estimate the green light time needed for crossing. It can also be used to count the number of pedestrians at the entrance or exit of public places, so as to control the population density. In addition, monitoring the number of people in some private places (such as hotel rooms) is also one of its applications. In this paper, the proposed infrared multi-pedestrian tracking algorithm combined with Faster-RCNN and SiamFC gives full play to the advantages of these two methods in their respective fields. Finally, a series of strategies are used to achieve multi-pedestrian tracking. At the same time, the similarity evaluation process of SiamFC is improved to increase the tracking accuracy during the change of pedestrian posture. The experimental results of infrared video recorded by the actual situation prove that the proposed algorithm is very effective and achieves competitive performance.

The tracking algorithm in this paper focus on the change of pedestrian posture and tracking multi-pedestrian without considering the challenges like occlusions, out-of-view, background cluttering and other variations. And the tracking accuracy would be greatly affected when these challenges occur. So, in the future work, we try to use higher quality training data to train a more robust detection and tracking model to deal with occlusions and complex backgrounds. Moreover, a re-detect strategy will be adopted when there is severe out-of-view and full occlusion.

REFERENCES

[1] G. Shen, C. Chen, Q. Pan, S. Shen, and Z. Liu, "Research on traffic speed prediction by temporal clustering analysis and convolutional neural network with deformable kernels," *IEEE Access*, vol. 6, pp. 51756–51765, 2018, doi: 10.1109/ACCESS.2018.2868735.
[2] G. Shen, X. Han, J. Zhou, Z. Ruan, and Q. Pan, "Research on intelligent analysis and depth fusion of multi-source traffic data," *IEEE Access*, vol. 61, pp. 59329–59335, 2018, doi: 10.1109/ACCESS.2018.2872805.
[3] X. Kong, M. Li, T. Tang, K. Tian, L. Moreira-Matias, and F. Xia, "Shared subway shuttle bus route planning based on transport data analytics," *IEEE Trans. Autom. Sci. Eng.*, vol. 15, no. 4, pp. 1507–1520, Oct. 2018.
[4] L. Wang, C. Li, M. Z. Q. Chen, Q.-G. Wang, and F. Tao, "Connectivity-based accessibility for public bicycle sharing systems," *IEEE Trans. Autom. Sci. Eng.*, vol. 15, no. 4, pp. 1521–1532, Oct. 2018.
[5] C. Sun, D. Wang, H. Lu, and M.-H. Yang, "Correlation tracking via joint discrimination and reliability learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 489–497.
[6] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: Efficient convolution operators for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6931–6939.
[7] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 3074–3082.
[8] Y. Song, C. Ma, L. Gong, J. Zhang, R. W. H. Lau, and M.-H. Yang, "CREST: Convolutional residual learning for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2574–2583.
[9] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4293–4302.
[10] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. S. Torr, "End-to-end representation learning for correlation filter based tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5000–5008.
[11] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4353–4361.

[12] A. He, C. Luo, X. Tian, and W. Zeng, "A twofold siamese network for real-time object tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Feb. 2018, pp. 4834–4843.
[13] D. Held, S. Thrun, and S. Savarese, "Learning to track at 100 FPS with deep regression networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 749–765.
[14] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-convolutional siamese networks for object tracking," in *Proc. Eur. Conf. Comput. Vis. Workshop*, 2016, pp. 850–865.
[15] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with siamese region proposal network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8971–8980.
[16] H. Zhang, W. Ni, W. Yan, J. Wu, H. Bian, and D. Xiang, "Visual tracking using Siamese convolutional neural network with region proposal and domain specific updating," *Neurocomputing*, vol. 275, pp. 2645–2655, Jan. 2018.
[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[18] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.
[19] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 779–788.
[20] K. Simonyan and A. Zisserman. (2014) "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556
[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[22] R. Tao, E. Gavves, and A. W. M. Smeulders, "Siamese instance search for tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1420–1429.
[23] R. Girshick. (2015). "Fast R-CNN." [Online]. Available: https://arxiv.org/abs/1504.08083
[24] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.
[25] J. Dai, Y. Li, K. He, and J. Sun. (2016). "R-FCN: Object detection via region-based fully convolutional networks." [Online]. Available: https://arxiv.org/abs/1605.06409
[26] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu, "Distractor-aware siamese networks for visual object tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 103–119.
[27] Q. Wang, Z. Teng, J. Xing, J. Gao, W. Hu, and S. Maybank, "Learning attentions: Residual attentional siamese network for high performance online visual tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4854–4863.

GUOJIANG SHEN received the B.Sc. degree in control theory and control engineering and the Ph.D. degree in control science and engineering from Zhejiang University, Hangzhou, China, in 1999 and 2004, respectively. He is currently a Professor with the College of Computer Science and Technology, Zhejiang University of Technology. His current research interests include artificial intelligence, theory, big data analytics, and intelligent transportation systems.

LINFENG ZHU received the bachelor's degree in computer science technology from the Zhejiang University of Technology, Hangzhou, China, in 2016, where he is currently pursuing the master's degree in computer technology. His current research interests include pedestrian detection and pedestrian tracking.

JIHAN LOU received the M.S. degree in automatic control theory and application from Beijing Jiaotong University, Beijing, China, in 1990. He is currently a Manager with Zhejiang Supcon Information Co., Ltd. His current research interests include electronic information communication and automatic control.

SI SHEN received the B.S. degree in computer science and technology from Luoyang Normal University, in 2011, and the M.S. degree from the Criminal Investigation Police University of China, Shenyang, China, in 2014. She is currently pursuing the Ph.D. degree with the College of Computer Science and Technology, Zhejiang University of Technology. Her current research interests include intelligent transportation and artificial intelligence.

ZHI LIU was born in Luoyang, Henan, China, in 1969. She received the B.S. degree in automatic control and the M.S. degree in system engineering from Xi'an Jiaotong University, Xi'an, China, in 1991 and 1994, respectively, and the Ph.D. degree in computer science and technology from Zhejiang University, Hangzhou, China, in 2001. She is currently a Professor with the College of Computer Science and Technology, Zhejiang University of Technology. She is a member of the China Computer Federation. Her current research interest includes intelligent transportation systems.

LONGFENG TANG received the bachelor's degree in computer science and technology from the Zhejiang University of Technology, Hangzhou, China, in 2017, where he is currently pursuing the master's degree in computer technology. His current research interests include intelligent transportation and artificial intelligence.